Robust bimodal person identification using face and speech with limited training data and corruption of both modalities

Robust Bimodal Person Identification using Face and Speech with Limited Training Data and Corruption of Both Modalities

Niall McLaughlin, Ji Ming, Danny Crookes

Institute of ECIT, Queen's University Belfast, Belfast, BT3 9DT, UK

nmclaughlin02@qub.ac.uk, j.ming@qub.ac.uk, d.crookes@qub.ac.uk

Abstract

This paper presents a novel method of audio-visual fusion for person identification where both the speech and facial modalities may be corrupted, and there is a lack of prior knowledge about the corruption. Furthermore, we assume there is a limited amount of training data for each modality (e.g., a short training speech segment and a single training facial image for each person). A new representation and a modified cosine similarity are introduced for combining and comparing bimodal features with limited training data as well as vastly differing data rates and feature sizes. Optimal feature selection and multicondition training are used to reduce the mismatch between training and testing, thereby making the system robust to unknown bimodal corruption. Experiments have been carried out on a bimodal data set created from the SPIDRE and AR databases with variable noise corruption of speech and occlusion in the face images. The new method has demonstrated improved recognition accuracy.

Index Terms: Multimodality, robustness, speaker recognition, face recognition, person identification

1. Introduction

In this paper we consider fusion of speech and facial data for person identification, assuming corruption of both modalities and limited training data, i.e. insufficient speech data to build a statistical model such as a GMM for each speaker, where there may be only one training facial image per person.

Commonly, multimodal fusion is performed at the feature or score level. Feature level fusion offers opportunities to optimise performance at an earlier stage, however it may be difficult in practice because of incompatible data rates and feature sizes between different modalities [1]. Interpolation may be used to combine speech, which typically has 100 frames per second, with video data with a much lower frame rate [2] [3]. We extend this approach to consider the combination of a single face image with a short speech segment.

If the feature vectors to be combined have very different scales, such as Gabor image features which typically have thousands of parameters, and the mel-frequency cepstral coefficients (MFCC) of speech which have a much smaller number of parameters, dimensionality reduction methods such as PCA or LDA may be used [4] [5]. However, given limited training data, it may not be feasible to perform such dimensionality reduction. Recent work has attempted to solve the problem of limited training data by adapting universal background models (UBM) [6] or by using fuzzy vector quantisation [7].

In this paper, we describe a novel method for fusion of speech and facial data for person identification, which simultaneously tackles all the four problems mentioned above, i.e., corruption of both modalities, limited training data for both modalities (e.g., a short training speech segment and a single training facial image), vast differences in the bimodal data rates, and in the feature sizes. We use a new representation to combine bimodal features while accommodating the different data rates, and a new similarity measure to compare the bimodal features, which enables differently sized feature vectors and limited training data. The system is further extended to include multi-style training and missing-feature theory, making it robust to noise corruption in speech and partial occlusion of the face while assuming minimal information about the corruption.

2. Similarity-based bimodal person recognition

We consider the combination of a short speech segment from person λ, of T frames $X^\lambda = (x^\lambda(1), x^\lambda(2), ..., x^\lambda(T))$, where $x^\lambda(t)$ is a frame vector at time t, and a single face image I^λ to form a model for the person. To accommodate corruption in either or both of the modalities, we represent each speech frame as F non-overlapped subbands, i.e., $x^\lambda(t) = (x_1^\lambda(t), x_2^\lambda(t), ..., x_F^\lambda(t))$ where $x_f^\lambda(t)$ is the feature for subband f in frame $\mathbf{x}^\lambda(t)$. Similarly, we represent each face image as K non-overlapped sub-images, i.e., $I^\lambda = (I_1^\lambda, I_2^\lambda, ..., I_K^\lambda)$ where I_k^λ represents the kth sub-image. To overcome the problem of differing data rates between the two modalities, we combine face image I^λ with every speech frame $x^\lambda(t)$ to form a new bimodal time sequence $\mathbf{X}^\lambda$

$$\begin{aligned}\mathbf{X}^\lambda &= \{(x^\lambda(1), I^\lambda), (x^\lambda(2), I^\lambda), ..., (x^\lambda(T), I^\lambda)\} \\ &= \{\mathbf{x}^\lambda(1), \mathbf{x}^\lambda(2), ..., \mathbf{x}^\lambda(T)\} \end{aligned} \quad (1)$$

In this new bimodal time sequence, each frame $\mathbf{x}^\lambda(t)$ groups together the speech subband features at time t and the whole face image represented by sub-images

$$\mathbf{x}^\lambda(t) = (x_1^\lambda(t), x_2^\lambda(t), ..., x_F^\lambda(t), I_1^\lambda, I_2^\lambda, ..., I_K^\lambda) \quad (2)$$

This representation allows for a static image to be combined with an arbitrary number of speech frames, eliminating the problem of differing data rates.

In recognition, let $Y = (y(1), y(2), ..., y(\Gamma))$ be a test speech segment of Γ frames and J be a test image, from an unknown person. In the same way as before, we represent each speech frame in subbands and the face image in sub-images, and form a bimodal time sequence for the person $\mathbf{Y} = \{\mathbf{y}(1), \mathbf{y}(2), ..., \mathbf{y}(\Gamma)\}$, where each bimodal frame $\mathbf{y}(t) = (y_1(t), y_2(t), ..., y_F(t), J_1, J_2, ..., J_K)$. Let $C(\mathbf{Y}, \mathbf{X}^\lambda)$ represent a similarity measure between the test sequence $\mathbf{Y}$ and a model sequence $\mathbf{X}^\lambda$ for person λ. We identify the unknown

person as follows, assuming text-independent training and test speech segments

$$\hat{\lambda} = \arg\max_{\lambda} C(\mathbf{Y}, \mathbf{X}^{\lambda}) = \arg\max_{\lambda} \sum_{t=1}^{\Gamma} \max_{\tau} C(\mathbf{y}(t), \mathbf{x}^{\lambda}(\tau)) \quad (3)$$

In order to perform text-independent speaker recognition, for each test frame $\mathbf{y}(t)$ we select the best matching model frame $\mathbf{x}^{\lambda}(\tau)$ for comparison.

Typically, a speech frame $x^{\lambda}(t)$ of 20 ms long can be represented using 30-40 features, covering 5-10 subbands (e.g., subband MFCC [8]), while a facial sub-image I_k, of 20×20 pixels for example, could be represented by more than 10^4 coefficients (e.g., Gabor features). Without proper normalization, such a huge disparity in feature sizes may cause the features from one modality to completely dominate in the comparison. In the following, we introduce a novel similarity measure, modified cosine similarity, for combining and comparing modalities of different sizes which effectively overcomes this problem.

The standard cosine similarity $C(\mathbf{a}, \mathbf{b})$ between two vectors $\mathbf{a} = (a_1, a_2, ..., a_Q)$ and $\mathbf{b} = (b_1, b_2, ..., b_Q)$, each composed of Q local vectors, can be expressed as

$$C(\mathbf{a}, \mathbf{b}) = \sum_{q=1}^{Q} \frac{a_q \cdot b_q}{||a_q||||b_q||} \frac{||a_q||||b_q||}{||\mathbf{a}||||\mathbf{b}||} = \sum_{q=1}^{Q} C(a_q, b_q) w_q \quad (4)$$

where $C(a_q, b_q) = a_q \cdot b_q / ||a_q||||b_q||$ is the inner product between local vectors a_q and b_q normalized by their respective norms. From (4) we can see that the overall cosine similarity is the sum of all the local cosine similarities $C(a_q, b_q)$ weighted by w_q, which equals the norms of the appropriate local vectors compared to the norms of the overall vectors. As the weight w_q is a function of the overall norms, it will be affected by any local corruption in either $\mathbf{a}$ or $\mathbf{b}$. In other words, the weighting can spread local vector corruptions globally. To avoid this problem, we assume a uniform weight w_q for all the local vectors, meaning they contribute equally to the overall similarity. Thus, we use a uniformly-weighted cosine similarity to compare the two multimodal frames, $\mathbf{y}(t)$ and $\mathbf{x}^{\lambda}(\tau)$, required in (3). This can be written as

$$C(\mathbf{y}(t), \mathbf{x}^{\lambda}(\tau)) \simeq \sum_{f=1}^{F} C(y_f(t), x_f^{\lambda}(\tau)) + \sum_{k=1}^{K} C(J_k, I_k^{\lambda}) \quad (5)$$

Note that, since each similarity measure $C(y_f(t), x_f^{\lambda}(\tau))$ of speech subbands and $C(J_k, I_k^{\lambda})$ of facial sub-images in (5) varies in the same range, from -1 to 1, all speech subbands and face sub-images contribute equally to the overall similarity, independent of the vast size disparity between the two local modality features.

Equation (5) can be expressed in an equivalent form

$$p(\mathbf{y}(t)|\mathbf{x}^{\lambda}(\tau)) = H^{C(\mathbf{y}(t), \mathbf{x}^{\lambda}(\tau))} = \prod_{f=1}^{F} H^{C(y_f(t), x_f^{\lambda}(\tau))} \prod_{k=1}^{K} H^{C(J_k, I_k^{\lambda})} \quad (6)$$

where $H > 1$ is a positive base number. The function $p(\mathbf{y}(t)|\mathbf{x}^{\lambda}(\tau))$ shares the characteristics of an exponent-type likelihood function for the test frame $\mathbf{y}(t)$ associated with speaker λ, given the model frame $\mathbf{x}^{\lambda}(\tau)$. Correspondingly, the recognition rule (3) can now be written as

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{t=1}^{\Gamma} \max_{\tau} \log p(\mathbf{y}(t)|\mathbf{x}^{\lambda}(\tau)) \quad (7)$$

3. Robustness to corruption

The system as currently defined assumes that both the training and test data of the speech and face image are uncorrupted. We extend the system to be resistant both to background noise for the speech modality and to partial occlusion for the facial modality. We achieve this by firstly modifying the computation of the likelihood $p(\mathbf{y}(t)|\mathbf{x}^{\lambda}(\tau))$ of a noisy bimodal test frame $\mathbf{y}(t)$ associated with a clean bimodal model frame $\mathbf{x}^{\lambda}(\tau)$, to incorporate multicondition training. Let $\mathbf{X}^{\lambda} = (\mathbf{x}^{\lambda}(1), .., \mathbf{x}^{\lambda}(T))$ be the given clean bimodal training sequence for λ, and $\mathbf{X}^{\lambda,l} = (\mathbf{x}^{\lambda,l}(1), .., \mathbf{x}^{\lambda,l}(T))$, $l = 1, 2, ..., L$, represent L multicondition training sequences generated from $\mathbf{X}^{\lambda}$, where each $\mathbf{X}^{\lambda,l}$ simulates a different corruption condition, with $\mathbf{X}^{\lambda,0}$ corresponding to the clean condition. These multicondition training sequences are combined to model a test bimodal sequence $\mathbf{Y}$ with feature corruption. The likelihood of a noisy test frame $\mathbf{y}(t)$ associated with a clean model frame given multicondition training can be written as

$$p(\mathbf{y}(t)|\mathbf{x}^{\lambda}(\tau)) = \sum_{l=0}^{L} p(\mathbf{y}(t)|\mathbf{x}^{\lambda,l}(\tau)) \quad (8)$$

where $p(\mathbf{y}(t)|\mathbf{x}^{\lambda,l}(\tau))$ is the likelihood of the noisy test frame $\mathbf{y}(t)$ associated with the model frame $\mathbf{x}^{\lambda}(\tau)$ corrupted at condition l. In our experiments, instead of assuming *a priori* knowledge about the test data corruption, for both speech and image, we try to compensate for a wide range of corruptions through multicondition training. For example, we add wide-band noise to the clean speech training data at different signal-to-noise rations (SNRs), to simulate a variety of unknown acoustic noises.

Secondly, in order to reduce mismatches between the actual and simulated corruption, we introduce optimal feature selection. At each corruption condition l, instead of using the full feature set $\mathbf{y}(t)$ to calculate the likelihood (i.e. (8)), we calculate the likelihood by choosing the subset of features in $\mathbf{y}(t)$ matched by the training condition l. This is based on missing-feature theory or the recognition-by-parts principle. Let $\mathbf{y}_{\mathbf{r}_l}(t) \subseteq \mathbf{y}(t)$ represent the matched, or reliable, test feature set in $\mathbf{y}(t)$ for training condition l, where $\mathbf{r}_l = \{f; k\}$, with $f \in (1, 2, ..., F)$ and $k \in (1, 2, ..., K)$, is an index set defining the optimal speech subband and image block features. Given a noisy test frame, we compute the optimal feature set $\mathbf{y}_{\mathbf{r}_l}(t)$ for each training condition by maximizing the probability of the corresponding model frame at that condition, i.e. $P(\mathbf{x}_{\mathbf{r}_l}^{\lambda,l}(\tau)|\mathbf{y}_{\mathbf{r}_l}(t))$. Using Bayes' theorem, this probability can be expressed as

$$P(\mathbf{x}_{\mathbf{r}_l}^{\lambda,l}(\tau)|\mathbf{y}_{\mathbf{r}_l}(t)) = \frac{p(\mathbf{y}_{\mathbf{r}_l}(t)|\mathbf{x}_{\mathbf{r}_l}^{\lambda,l}(\tau))}{\sum_{\lambda'} \sum_{l'} \sum_{\tau'} p(\mathbf{y}_{\mathbf{r}_l}(t)|\mathbf{x}_{\mathbf{r}_l}^{\lambda',l'}(\tau')) + \epsilon} \quad (9)$$

The sum in the denominator is over the model feature sets of all the frames, taking into account all the training speakers and conditions (including the clean training condition), and ϵ is a small positive number accounting for any test set $\mathbf{y}_{\mathbf{r}_l}(t)$ without matching model sets $\mathbf{x}_{\mathbf{r}_l}(\tau)$ (hence the sum approaches zero).

We can obtain an optimal estimate for the unknown index set $\mathbf{r}_l$ at each training condition l, by maximizing the posterior probability (9). Therefore the recognition rule, combining both multicondition training and optimal feature estimation, can be expressed as

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{t=1}^{\Gamma} \max_{\tau} \log \{ \sum_{l=0}^{L} \max_{r_l} P(\mathbf{x}_{r_l}^{\lambda,l}(\tau)|\mathbf{y}_{r_l}(t)) \} \tag{10}$$

In (10) the optimal feature index set for each test frame is estimated at each training condition, and the contributions of all the training conditions are summed towards the overall similarity. In this paper we demonstrate that the above system (10) can be made robust to unknown full band corruption in the speech modality and partial occlusion in the facial modality. As pointed out earlier it is capable of accommodating the vastly different data rates and feature sizes between speech and facial modalities. In addition the system can be used with very limited training data in both modalities.

4. Experiments

We first compare the resistance of our system against an oracle model for band-limited noise corruption, using the speech modality alone. The oracle model used prior knowledge of the corruption to remove the affected subbands before performing recognition, hence acting as a type of "idealised" noise robust system. Our system did not have knowledge of the noise corruption. This experiment was carried out using test samples from the SPIDRE speaker database, which contains 45 speakers. Each speech sample was silence-stripped and divided into 20 ms frames overlapping by 10 ms. Each frame was processed through a 22-channel log mel-scale filter and the filter outputs decorrelated with a high pass filter, giving 21 decorrelated log mel filter bank coefficients. These coefficients were uniformly placed into groups of three, giving seven subband features. First-order derivative coefficients were included, resulting in 14-subband feature streams for each frame, each stream containing three elements. The value of H used in these experiments was 15000 and the value of ϵ was 10^{-5}. In this experiment the system was trained using 30 s of clean speech and tested using five 10 s samples of band-corrupted speech from each speaker. Test-samples from each speaker were corrupted using 0 dB band-limited noise at various centre frequencies and bandwidths. The 'do-nothing' system, which did not perform any noise compensation, serves as a baseline. From the results in Table 1 we see that for the most part our proposed system performed better than the oracle model and always significantly better than the do-nothing model. This indicates that the system is capable of removing the contribution of noise corrupted speech subbands from each speaker's score. The fact that our system could outperform the oracle model in many cases, may be due to the fact that the oracle model removes all bands believed to be corrupted. It may be the case that some features e.g. the delta features of some corrupted bands, are only partially corrupt and thus are still usable for recognition. This indicates that a softer masking approach may be preferable to a binary masking approach.

Table 1: *Speaker identification accuracy (%) with subband corruption, comparing our system to an oracle model and 'do-nothing' baseline, for different noise central frequencies, bandwidths and affected subbands (out of seven).*

Corruption Properties			Accuracy (%)		
Centre (Hz)	B/width (Hz)	Noisy Bands	Our System	Oracle	Do Nothing
656	175	1	68	65.3	58.7
1031	225	2	64	60	51.6
1265	325	3	46.2	45.3	28.9
2156	400	3	47.6	48	28.9

Secondly, a speaker identification experiment was performed using noise corrupted test samples created by adding realistic non-stationary, full-band noise at 10 dB, 15 dB and 20 dB to each test sample from the SPIDRE database (See Fig. 1). The noise types used were pop-song, restaurant and street noise. Noisy speaker models were created using multi-condition training by adding low-pass filtered white noise, with a 3 dB cut-off frequency of 2 kHz, at SNRs from 10 dB to 20 dB in 5dB steps to each clean training segment. In addition to tests of the proposed system, tests were performed with the system using multicondition training only, optimal feature selection only and 'do-nothing' model. The results are presented in Table 3. We can see that our system significantly outperformed the do-nothing model at all the low SNR conditions for the untrained realistic full-band noise, with some performance loss at a few clean or higher SNR conditions. We also see that multicondition training and optimal feature selection make independent contributions to overall system performance.

An experiment in person identification using the facial modality alone was performed on the AR face database which consists of the frontal face images of 126 persons with realistic partial occlusions. A random selection of 45 persons was used for testing. Each face was downsized to 59×43 pixels, and split into 16 equal size sub-images. Each sub-image was processed using the gradient face algorithm [9] before Gabor features were extracted at 4 scales and 4 orientations, then concatenated to form a 2240 element feature vector. Thus, each face was represented by a total of $2240 \times 16 = 35840$ elements. In these experiments, a single clean face image, randomly selected from the training set, was used as the training image for each person. Tests were carried out using three clean face images and three occluded face images from each occlusion condition – sunglasses and scarf – for each person (See Fig. 2).

Table 2: *Facial identification accuracy (%) with the AR database with a single training image by the proposed system.*

	Clean	Sunglasses	Scarf
Accuracy (%)	97.7	89.6	94.0

The results are shown in Table 2, which exceeded those of previously published results e.g. [10], (77% sunglasses, 89% scarf) on the same database using one training image per person.

Finally, bimodal experiments were performed, where either a single or both modalities were corrupted to varying degrees. For these experiments, each person was trained with a single clean face image, which was combined with simulated, multicondition training speech data derived from either 5 s or 10 s of clean training speech data as described earlier. Testing was performed using 3 speech samples of 10 s from each noise condition (street, restaurant, pop song), paired with a randomly chosen face image from one of the three facial occlusion conditions (clean, sunglasses and scarf) for each person. Our system was compared to a 'do-nothing' model, which combined all the components from both modalities, without performing

Table 3: *Speaker identification accuracy (%) tested on the SPIDRE speaker database with various realistic noise types added at variable SNRs. Tests were performed with both 5 s and 10 s of training data and 10 testing samples per speaker.*

		Clean	Restaurant			Street			Pop-song		
System	Training (s)		10dB	15dB	20dB	10dB	15dB	20dB	10dB	15dB	20dB
Our System (OS)	5	79.3	52.7	67.1	72.4	62.2	73.1	74.9	66.2	73.6	72.4
	10	81.5	63.6	73.3	76.0	70.9	77.3	78.4	73.6	76.7	78.2
Optimal Feature Only	5	82.2	39.3	55.6	68.1	41.5	55.6	69.6	57.1	72.2	76.2
	10	82.2	45.1	62.2	74	42.9	62.2	75.5	67.3	76.0	78.9
Multicondition Only	5	72.6	41.5	51.8	63.7	55.5	68.1	74	59.8	66.4	68.9
	10	81.5	51.9	66.7	77	65.9	74.8	82.2	67.6	77.1	80.2
Do-Nothing (DN)	5	80	31.9	47.7	60.7	39.3	51.9	68.8	56.3	68.2	73.5
	10	83.7	42.2	57.7	71.1	42.2	61.5	76.2	64.2	75.5	79.3

noise compensation. The results are summarised in Table 4. We

Table 4: *Bimodal person identification accuracy (%) with limited training speech and a single training facial image, using noisy test speech and occluded test images, comparing Our System (OS) and 'Do-Nothing' (DN). Note that the results for the clean face condition were all 100% and so have been omitted.*

Facial Occlusion/ Acoustic Noise	Train (s)	SNR (dB)	Sunglasses OS	Sunglasses DN	Scarf OS	Scarf DN
Clean	10		97	93.3	98.5	95.5
	5		96.3	94.1	97.8	94.1
Restaurant	10	10	97	93.3	97	98.5
		20	99.3	98.5	97.8	96.2
	5	10	96.3	92.5	97.8	96.2
		20	99.3	96.3	97.8	97.7
Street	10	10	97	92.6	97	95.6
		20	99.3	97	98.5	95.6
	5	10	95.6	92.6	97	95.6
		20	99.3	96.3	97.8	96.3
Pop-song	10	10	97.8	92.6	97.8	96.3
		20	98.5	97.7	98.5	97
	5	10	95.6	92.6	97	96.3
		20	97.8	96.3	97.8	97.8

can see that when both modalities are corrupted, the identification accuracy remains higher than the best accuracy achieved by either unimodal system tested on corresponding corruption conditions. For example, in the case of 10 dB restaurant noise with 10 s training and face occluded by sunglasses, the unimodal speaker and facial identification systems score, 63.6% and 89.6% respectively, while the bimodal system scores 97% given identical corruption conditions. In addition, our system frequently outperforms the 'do-nothing' model. These results demonstrate that by choosing the best uncorrupted sub-bands and sub-images from each person, our system is capable of improving identification accuracy compared to the component unimodal systems, as well as improving accuracy compared to the combination of all components from both modalities.

5. Conclusions

In this paper we have proposed a new method of bimodal person identification that can be used with limited training data and is robust to corruption in both modalities. We have shown that a modified cosine similarity can be used for comparing speech as well as bimodal feature vectors. We have also shown that it is possible to combine speech with a single face image at the feature level, despite the vast differences in data rate and feature size. Experiments have been performed which demonstrate the proposed method of bimodal combination shows improved person identification accuracy, compared to using unimodal features alone, even when both modalities are corrupted.

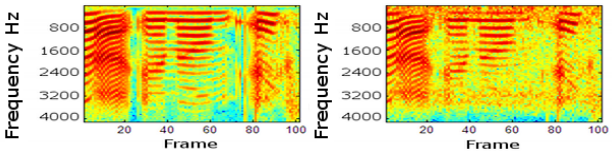

Figure 1: Spectrograms showing clean speech on the left, and the same sample corruption by 10 dB engine noise on the right.

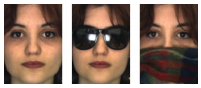

Figure 2: A sample face from the AR database showing the clean condition and realistic corruption by sunglasses and scarf.

6. References

[1] D. L. Hall, *Mathematical Techniques in Multisensor Data Fusion*. Artech House, 1992.

[2] S. M. O. C. Bregler and Y. Konig, "A hybrid approach to bimodal speech recognition," *Asilomar Conference on Signals, Systems and Computers*, vol. 1, p. 556, 1994.

[3] G. Potamianos and H. Graf, "Discriminative training of hmm stream exponents for audio-visual speech recognition," *ICASSP'98*, vol. 6, pp. 3733 – 3736, 1998.

[4] A. Ross and R. Govindarajan, "Feature level fusion of hand and face biometrics," *SPIE*, vol. 5779, pp. 196–204, 2005.

[5] J. M. C.C. Chibelushi and F. Deravi, "Feature-level data fusion for bimodal person recognition," *Image Processing and Its Applications, 1997.,*, vol. 1, pp. 399 – 403, 1997.

[6] P. Angkititrakul and J. H. L. Hansen, "Discriminative in-set/out-of-set speaker recognition," *Audio, Speech, and Language Processing*, vol. 15, p. 498, 2007.

[7] H. S. Jayanna and S. R. M. Prasanna, "Fuzzy vector quantization for speaker recognition under limited data conditions," *TENCON 2008*, pp. 1–4, 2009.

[8] C. J. L. Lin and S. Xiaoying, "A discriminative method for speaker identification with limited data," *(FSKD)*, vol. 2, p. 512, 2010.

[9] B. F. Z. S. T. Zhang, Y.Y. Tang and X. Liu, "Face recognition under varying illumination using gradientfaces," *Image Processing,*, vol. 18, p. 2599, 2009.

[10] J. I. L. Zisheng and M. Kaneko, "Robust face recognition using block-based bag of words," *(ICPR), 2010*, p. 1285, 2010.